# Sample Size and Item Parameter Estimation Precision When Utilizing the Masters' Partial Credit Model

Michael Custer and Jongpil Kim
Riverside Insights

Paper Presented at the Annual Meeting of the
National Council of Measurement in Education
Chicago, Illinois
March 28-30
April 12-15
2023

Please send correspondence regarding this paper to michael.custer@riversideinsights.com

## Abstract

This study utilizes an analysis of diminishing returns to examine the relationship between sample size and item parameter estimation precision when utilizing the Masters' Partial Credit Model for polytomous items.

Item data from the standardization of the Batelle Developmental Inventory, 3rd Edition were used. Each item was scored with a "0" for no credit, a "1" for partial credit and a "2" for full credit. Two conditions were studied; the first used 40 items and the second 20 items. Item parameter estimates were examined relative to their "true" values by evaluating the decline in root mean squared error (RMSE) and the number of outliers as the level of sample size increased.

The generalizability of these results may be limited due to this study's range of item score points, the number of items, and the range of the underlying scale. However, under both conditions, the majority of RMSE and outlier-reduction improvement was achieved once a sample size of 900 examinees had been reached. Diminishing returns were most notable with incremental increases to sample size beyond 1,000 examinees.

Practitioners often encounter great variability in the number of items, item types, the number of item score points, the range of an assessment's underlying scale and the stakes associated with calibration objectives. Given this variability, practitioners might be encouraged to first simulate item data and vary the level of sample size to evaluate estimation precision across the scale through a similar review of RMSE and "outliers" as the level of sample size increases. Besides the consideration of sample acquisition costs, an evaluation of "diminishing returns" could aid practitioners in their selection of an appropriate sample size.

## Background

There exists a positive relationship between incremental increases in sample size and estimation precision when utilizing both Rasch and Item Response Theory (IRT) models to estimate item parameters. An inadequate sample size can lead to increased estimation error with negative implications for the analysis of item and test data and IRT based test construction (Hambleton, Jones & Rogers, 1993; Swaminathan et al, 2003; He & Wheadon, 2012).

The Masters' Partial Credit Model (MPCM) (Masters, 1982) is a polytomous Rasch measurement model for items that have a set of ordered response categories for which credit is given for partially correct responses. Ordered response categories imply an increased ability on the trait being measured. For each item, "step" thresholds are estimated between adjacent response categories for which the probability of the two adjacent categories is equally likely. Given an examinee's ability, the "step" thresholds are used to estimate the probability of category response (Wright & Mok, 2000). The MPCM accommodates items with different response categories, a different number of response categories as well as different response formats (Mead, 2008).

Previous research has found that for polytomous items, larger sample sizes result in improved item parameter recovery. The increased estimation precision is the result of larger n-counts at each item score category which in turn provide more information for the estimation of item parameters (DeMars, 2003; Wollack, et al., 2002; De Ayala & Sava-Bolesta, 1999; Lee, 1997; Reise & Yu, 1990). Though these studies indicate that sample size is an important factor when utilizing Polytomous Rasch and Polytomous IRT models, there is little consensus regarding the minimum sample size needed for accurate estimation precision (Dai, et al., 2021). Reise and Yu (1990) evaluated estimation precision with the Graded Response Model and

recommended a minimum sample size of 500. DeAyala and Sava-Bolesta (1999) and DeMars (2003) studied the relationship between sample size and estimation precision using the Nominal Response Model. Both studies found that estimation precision improved with larger sample sizes. De Ayala and Sava-Bolesta (1999) suggested that the ratio of sample size to the number of model parameters could be used to identify an appropriate level of sample size.

Embretson and Reise (2000) recommended that researchers focus on the standard error of parameter estimates to ensure their reasonability. Wasserman and Bracken (2003) later noted that standard errors became smaller as sample size increased up to a point after which diminishing returns were expected. This work was further extended by He and Wheadon (2012) who utilized the MPCM and presented evidence that showed an association between increasing levels of sample size and a decline in item standard error estimates and the Root Mean Squared Error (RMSE) between item parameter estimates and their "true" values. Given the range of assessment stakes and item calibration goals, they recommended that practitioners select a sample size that yields a suitable RMSE for the calibration purpose.

## Method

Item data from the standardization of the Batelle Developmental Inventory, 3rd Edition (Newborg, 2020) (BDI-3) were used for this study. The BDI-3 measures developmental skills across five domains: Adaptive, Social-Emotional, Communication, Motor and Cognitive. The nationally representative standardization sample was comprised of 2,598 children aged between birth and 7 years 11 months old. The BDI-3 is an individually administered assessment with three item formats: structured, observation and parent-interview. Each item is scored with a "0" for no credit, a "1" for partial credit and a "2" for full credit. Item data for 40 items administered as part of the Expressive Communication subdomain were used for this study.

Two conditions were studied; in the first condition all 40 items were used and in the second condition, odd numbered items from the original set of 40 items were selected to derive a 20-item set. For the 40 item-set condition, “true” item parameter values were derived by calibrating the 40-item set with the full set of 2,598 examinees. To evaluate the relationship between sample size and item parameter recovery across 15 levels of sample size (15 levels ranging from 100 to 1,500 examinees in increments of 100), random samples of 100 through 1,500 cases were drawn from the 2,598 examinees. Ten replications of the random selection process were executed for each of the 15 levels of sample size. This process created 150 “40-item set” condition data sets (15 levels of sample size x 10 replications) to be used for the item calibrations. For the 20-item set condition, “true” item parameter values were derived by calibrating the 20-item set with the full set of 2,598 examinees. The same process was followed to derive the 150 “20-item set” condition data sets as was used to derive the 150 “40-item set” condition data sets. WINSTEPS 5.1.1.0 (Linacre, 2022) was used to calibrate each of the two “true” data sets as well as each of the 300 “condition” data sets with the MPCM. For each calibration the scale was centered to have an item mean of 0.00 and the value of one logit equal to 1.00.

The item difficulty measure reported by WINSTEPS was used to evaluate item parameter recovery. For a three point item with a maximum score of ‘2’, the reported item difficulty measure represents the location on the scale of the highest probability of a score of “1”. For each item, the highest probability of a score of “1” is the approximate midpoint between the measure associated with an item score of 0.5 (transition point between a score of 0 and 1) and the measure associated with a score of 1.5 (the transition point between a score of 1 and 2).

The RMSE for each item calibration was computed by summing the squared differences between the estimated and "true" measures and dividing this value by the number of items and taking the square root. The number of outlier items for each item calibration was defined as the number of item difficulty measures that differed from "true" by .30 or more. The threshold value of .30 reflects the notion that item difficulty measures that differ from their base calibration measure by less than .20 to .30 of a logit have no practical impact on person measurement (Wright, 1977). Also, this threshold was more than 4 times the average of the base calibration item standard errors of estimate which was equal to .071. For each level of sample size, the average RMSE and average number of outlier items was computed across the 10 runs.

## Results

For each condition, descriptive statistics for the "true" item parameter values are presented in Table 1. The item mean for each condition was equal to 0.00. For the 40-item set, the standard deviation (SD) and range of the item measures was respectively 5.992 and 23.427. The SD and range for the 20-item set was respectively 6.574 and 26.495. This high degree of variability is not surprising given the range in difficulty of items that are targeted to children ranging in age from birth through 7 years and 11 months. When comparing the 40-item set and 20-item set conditions, the 20-item set condition had a slightly larger SD and range, respectively 9.7% and 13.1%.

Descriptive statistics for persons are presented in Table 2. The ability mean for the 20-item set condition (1.746) was higher than that for the 40-item set condition (1.255). When comparing the 40-item and 20-item set conditions, the SD for person abilities with the 20-item set (7.299) was slightly larger than that for the 40-item set (6.597).

For both the 40-item set and 20-item set conditions, Table 3 presents the average RMSE for each level of sample size. Figure 1 presents the relationship between sample size and the average RMSE for the 40-item set condition and Figure 2 for the 20-item set condition. Likewise, Table 4 presents the average number of outliers for each level of sample size. Figure 3 presents the relationship between sample size and the average number of outliers for the 40-item set condition and Figure 4 for the 20-item set condition.

With the 40-item set condition, the decline in RMSE as sample size increases is presented in Table 3 and Figure 1. The majority of RMSE decline is reached by a sample size of 900 (RMSE=.216) and the shape of the curve begins to flatten at a sample size of 1,000 (RMSE = 0.178). This flattening effect, which is reflective of diminishing returns, is especially noticeable in Figure 3 which plots the number of outlier items. As shown in Table 4, at a sample size of 900 the number of outlier items is 10.50% and at a sample size of 1,000, 6.75%.

With the 20-item set condition, the decline in RMSE as sample size increases is depicted in Table 3 and Figure 2. The majority of RMSE decline is reached by a sample size of 900 (RMSE = 0.307). The improvement in outlier reduction is depicted in Figure 4. In a similar manner, the majority of outlier reduction improvement appears to be reached at a sample size of 900. As presented in Table 4, at a sample size of 900 the number of outlier items is 17.5%. However, outlier-reduction appears to improve to ~11.8% at sample sizes between 1,200 and 1,400 examinees.

## Discussion

Calibration objectives are important when practitioners consider the issue of estimation precision and sample size. For example, a high degree of estimation precision is typically

required for purposes such as scale construction or the calibration of items for an item bank. In contrast, less estimation precision is typically required for a pilot study item tryout.

Under the fairly unique conditions imposed by the study design, the results indicate that with both the 40-item set and 20-item set conditions, the majority of RMSE and outlier-reduction improvement was achieved once a sample size of 900 examinees had been reached. Diminishing returns were most notable with incremental increases to sample size beyond 1,000 examinees.

Practitioners often encounter great variability in the number of items, item types, the number of item score points, the range of an assessment's underlying scale and the stakes associated with calibration objectives. Given this variability, practitioners might be encouraged to first simulate item data and vary the level of sample size to evaluate estimation precision across the scale through a similar review of RMSE and "outliers" as the level of sample size increases. Besides the consideration of sample acquisition costs, an evaluation of "diminishing returns" could aid practitioners in their selection of an appropriate sample size.

## Suggestions for Future Research

The study of "diminishing returns" as it relates to sample size and item parameter estimation precision might be extended through the calibration and evaluation of simulated data across different levels of scale range. This might help answer the question; when the number of items is held constant, does a scale with a wide range in item difficulty require a larger sample size than a scale with a narrower range? In a similar manner, a mix of item formats and item score points could also be studied. Likewise, researchers might also be encouraged to evaluate the relationship between sample size and estimation precision across different levels of missing data or across different ability distributions when holding the number of items constant.

## References


Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement, 1*, 114-142.

Dai, S., Vo, T., Kehinde, O., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data, *Frontiers in Education, Vol. 6*, Article 721963.

De Ayala, R. J. & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

DeMars, C. E. (2003). Sample Size and the Recovery of Nominal Response Model Item Parameters. *Applied Psychological Measurement, 27(4),* 275-288

Embretson, S.E. & Reise, S. P. (2000). *Item Response Theory For Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associate.

Hambleton, R., Jones, R & Rogers, H. (1993). Influence of Item Parameter Estimation Errors in Test Development. *Journal of Educational Measurement, 30*, 143-155.

He, O & Wheadon, C. (2012). *The Effect of Sample Size on Item Parameter Estimation For the Partial Credit Model*. Centre for Education Research and Policy, www.cerp.or.uk

Lee, Y. S. (1997). *A parameter recovery study for the nominal response model*. Unpublished manuscript, Department of Educational Psychology, University of Wisconsin–Madison.

Linacre, J. M. (2022). *Winsteps® Rasch measurement computer program* (Version 5.1.1). Portland, Oregon: Winsteps.com

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mead, R. J. (2008). *A Rasch Primer: the measurement theory of Georg Rasch*. Psychometrics services research memorandum 2008-001. Maple Grove, MN: Data Recognition Corporation.

Newborg, J. (2020). *Battelle Developmental Inventory, 3rd Edition*. Riverside Assessments, LLC.

Reise, S. P. & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using Multilog. *Journal of Educational Measurement, 27*, 133-144.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D. & Rizavi, S. (2003). Small Sample Estimation in Dichotomous Item Response Models: Effect of priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates. *Applied Psychological Measurement, 27*, 27-51.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric Considerations of Assessment Procedures. In J. Graham and J. Naglieri (Eds*). Handbook of Assessment Psychology (pp. 43–66).* New York: Wiley.

Wollock, J., Bolt, D., Cohen., A., & Lee, Y. (2002). Recovery of Item Parameters in the Nominal Response Model: A Comparison of Marginal Maximum Likelihood Estimation and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement, 26(3),* 339–352.

Wright, B. & Mok, M. (2000). Rasch Models Overview. *Journal of Applied Measurement, 1(1),* 83-106.

Wright, B. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement, 14 (2),* pp. 97-116, Summer 1977 (and MESA Memo 42)

**Table 1. Descriptive Statistics for Items: "True Parameters"**

| Data Set | Item Difficulty Mean | Item Difficulty SD | Item Difficulty Minimum | Item Difficulty Maximum |
|---|---|---|---|---|
| **40 Items** | 0.000 | 5.992 | -13.788 | 9.639 |
| **20 Items** | 0.000 | 6.574 | -15.692 | 10.803 |

**Table 2. Descriptive Statistics for Persons: "True Parameters"**

| Data Set | Raw Score Mean | Raw Score Mean / n-Items | Raw Score SD | Ability Mean | Ability SD | Ability Minimum | Ability Maximum |
|---|---|---|---|---|---|---|---|
| **40 Items** | 45.864 | .573 | 25.293 | 1.255 | 6.597 | -15.465 | 12.407 |
| **20 Items** | 23.354 | .584 | 12.786 | 1.746 | 7.299 | -18.186 | 13.643 |

**Table 3. RMSE and Change as Sample Size Increases for the 40 Item and 20 Item Sets**

| Sample Size | 40 Items | | 20 Items | |
|---|---|---|---|---|
| | Item RMSE | Diff from Prev. Level | Item RMSE | Diff from Prev. Level |
| **100** | 0.803 | | 1.002 | |
| **200** | 0.677 | -0.125 | 0.676 | -0.326 |
| **300** | 0.652 | -0.026 | 0.772 | 0.096 |
| **400** | 0.442 | -0.209 | 0.536 | -0.236 |
| **500** | 0.363 | -0.079 | 0.442 | -0.093 |
| **600** | 0.301 | -0.062 | 0.446 | 0.004 |
| **700** | 0.265 | -0.036 | 0.386 | -0.060 |
| **800** | 0.260 | -0.005 | 0.429 | 0.043 |
| **900** | 0.216 | -0.044 | 0.307 | -0.122 |
| **1,000** | 0.178 | -0.037 | 0.369 | 0.061 |
| **1,100** | 0.260 | 0.082 | 0.352 | -0.017 |
| **1,200** | 0.156 | -0.104 | 0.289 | -0.062 |
| **1,300** | 0.150 | -0.006 | 0.286 | -0.004 |
| **1,400** | 0.139 | -0.011 | 0.255 | -0.030 |
| **1500** | 0.154 | 0.015 | 0.404 | 0.149 |

**Figure 1. 40 Item Set: The Reduction in RMSE as Sample Size Increases**

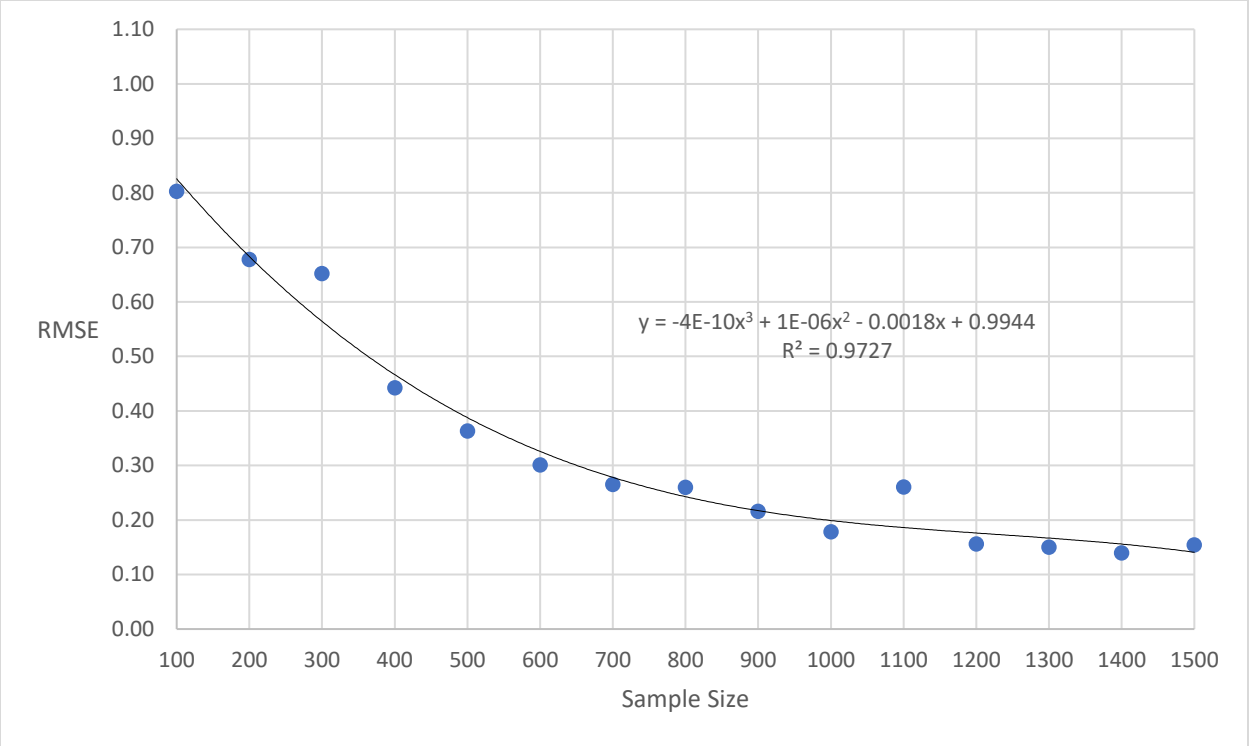


**Figure 2. 20 Item Set: The Reduction in RMSE as Sample Size Increases**

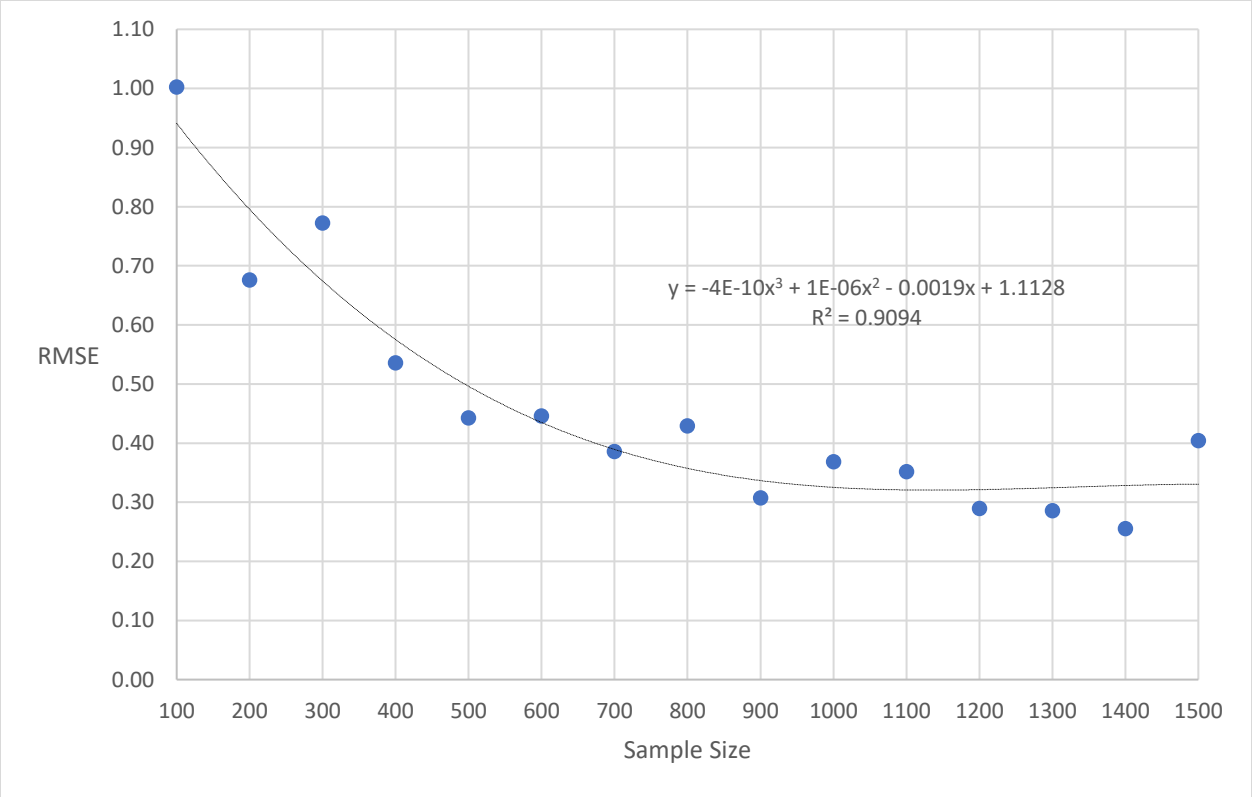

**Table 4. The Number of Items with Item Difficulty Measures that Differ From "True Parameter Value" By .30 or More As Sample Size Increases**

| Sample Size | 40 Items | | 20 Items | |
|---|---|---|---|---|
| | # of Items w/ Diff ≥ .30 | Percent n / 40 | # of Items w/ Diff ≥ .30 | Percent n / 20 |
| **100** | 26.30 | 65.75% | 12.80 | 64.00% |
| **200** | 25.00 | 62.50% | 13.00 | 65.00% |
| **300** | 20.90 | 52.25% | 12.20 | 61.00% |
| **400** | 12.90 | 32.25% | 8.60 | 43.00% |
| **500** | 12.30 | 30.75% | 8.70 | 43.50% |
| **600** | 8.30 | 20.75% | 6.10 | 30.50% |
| **700** | 8.70 | 21.75% | 5.10 | 25.50% |
| **800** | 7.30 | 18.25% | 6.30 | 31.50% |
| **900** | 4.20 | 10.50% | 3.50 | 17.50% |
| **1,000** | 2.70 | 6.75% | 4.50 | 22.50% |
| **1,100** | 5.30 | 13.25% | 4.30 | 21.50% |
| **1,200** | 2.30 | 5.75% | 2.90 | 14.50% |
| **1,300** | 2.40 | 6.00% | 1.80 | 9.00% |
| **1,400** | 1.40 | 3.50% | 2.40 | 12.00% |
| **1,500** | 2.20 | 5.50% | 2.60 | 13.00% |

**Figure 3. 40 Item Set: The Reduction in Number of Outliers as Sample Size Increases**

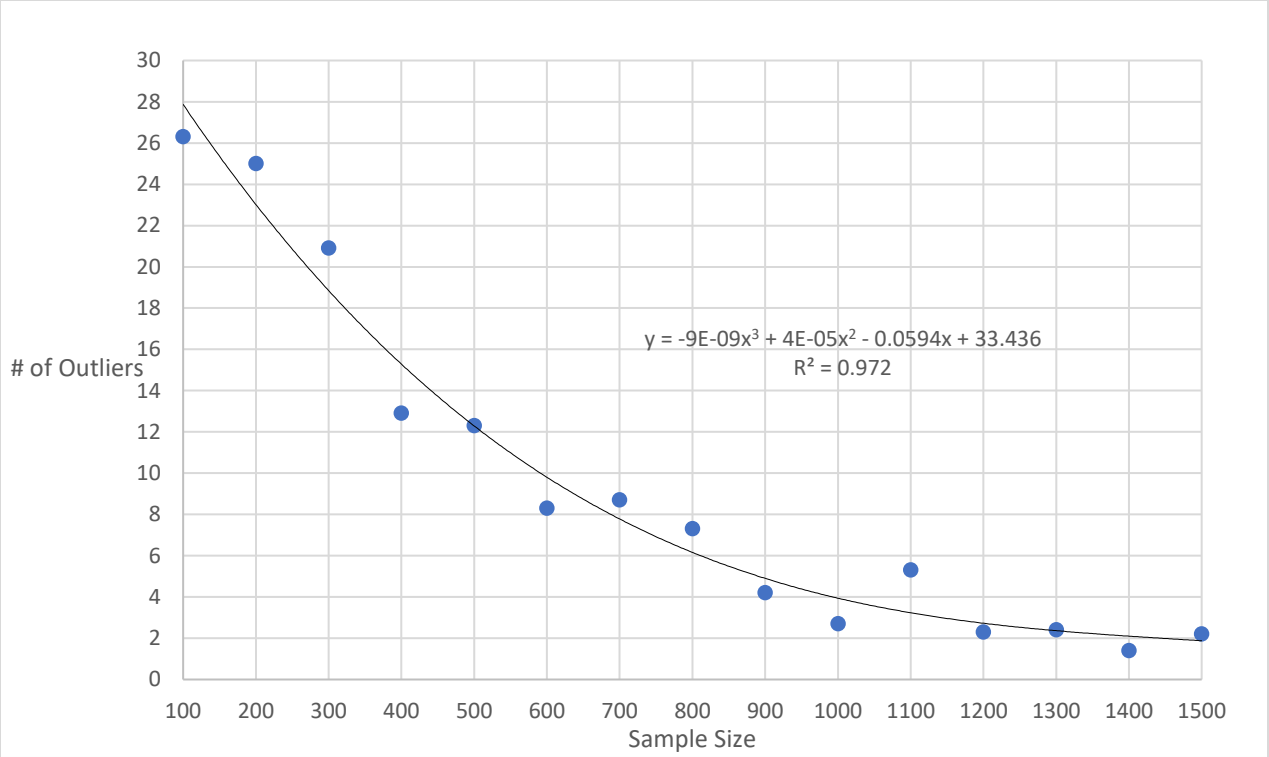


**Figure 4. 20 Item Set: The Reduction in Number of Outliers as Sample Size Increases**